

DOCUMENT RESUME

ED 306 240

TM 013 003

AUTHOR Huntley, Renee M.; And Others
TITLE Do Verbal Factors Affect Performance in Mathematics Tests?
PUB DATE Mar 89
NOTE 17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, March 28-30, 1989).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; College Entrance Examinations; High Schools; *High School Students; Item Analysis; *Mathematics Tests; Reading Skills; Sex Differences; Test Construction; *Test Format; Test Results; *Verbal Ability; Verbal Tests
IDENTIFIERS American College Testing Program

ABSTRACT

This study investigated the effects of verbal factors on item performance in mathematics tests. Experimental tryout units were constructed that varied the verbal load of each of 21 mathematics test items. Each tryout unit was randomly assembled to include A, B and C versions of items so that no examinee took a unit of only one item format. Each unit was administered to approximately 3,000 examinees as part of the American College Testing tryout procedure. Results suggest that increasing the verbal load of an item does affect item performance; as the verbal load in mathematics tests increases, the difficulty of the items increases. It may be that increasing the verbal load necessitates the use of verbal and reading skills as well as mathematics skills. Results of both analyses of variance and the standardization procedure indicate that there is little difference in performance between males and females on the types of items studied as verbal load increased. Four tables present study data. An appendix shows the three versions of a sample item. (SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED306240

Do Verbal Factors Affect Performance in Mathematics Tests?

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

RENEE M. HUNTLEY

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Renee M. Huntley

American College Testing Program

Terry A. Ackerman

American College Testing Program

Catherine Welch

American College Testing Program

A paper presented at the annual meeting of the
National Council on Measurement in Education
San Francisco, March 1989

TM013003

Do Verbal Factors Affect Performance
in Mathematics Tests?

Renee M. Huntley
American College Testing Program

Terry A. Ackerman
American College Testing Program

Catherine Welch
American College Testing Program

ABSTRACT

This study investigated the effect of verbal factors on item performance in mathematics tests. Experimental tryout units were constructed that varied the verbal load of each of twenty-one mathematics test items. Results suggest that, as the verbal load in mathematics items increases, the items become more difficult. However, unlike previous studies, this study found no statistically significant difference between male and female examinees as the degree of verbal loading increased.

INTRODUCTION

The purpose of this study is to investigate the effect of verbal factors on item performance in mathematics tests. It is generally conceded that test scores may be affected by reading ability even when the ostensible goal of a particular test may be to measure an ability other than reading ability (Aiken, 1971a; Muscio, 1962). This relationship, however, has been less clear in mathematics tests than in other tests (Paul, Nibbelink, & Hoover, 1986). In a mathematics test, for example, reading may not be a necessary component of the construct purported to be tested but, depending upon the format in which the test item is cast, reading or verbal skills may intrude sufficiently to raise questions about the meaning of a test score. That is, would an examinee's seemingly poor performance signal a lack of knowledge in the mathematical subject area, or would it indicate some verbal difficulties or lack of reading skills?

Furthermore, verbal factors in mathematics items might affect examinee performance differentially by gender. Although Benbow and Stanley (1982) reported that male high school students generally outperform female high school students in mathematics, Doolittle and Cleary (1987), who analyzed several forms of the American College Testing Program's Mathematics Usage Test, found that males outperformed females on items that were classified as Algebraic and Arithmetic Reasoning (AAR) items. These items were described as being "word problems," and thus could be viewed as representative of items that have a large reading or verbal component.

In their study of verbal and computational problems in mathematics, Caldwell and Golden (1979) described such items by yet another classification scheme: (1) Abstract versus Concrete--a problem about digits in a number is considered abstract, while one about baseballs is considered concrete; and

(2) Factual versus Hypothetical--a problem that supposes a possible change in the situation it has just described is considered to be hypothetical (and is invariably introduced by an "If" clause). For elementary school children, Caldwell and Golden found that the concrete problems (about baseballs) were substantially less difficult than the abstract ones (about digits), and "female students solved somewhat more abstract problems and fewer concrete problems than did male students."

The present study will explore the following questions: 1) Does varying the verbal load of mathematics items affect student performance? 2) As the verbal load is increased in mathematics items, will the performance of male examinees significantly differ from that of female examinees?

METHOD

Three versions were constructed for each of twenty-one mathematics items such that Version A of the item was a pure mathematical-operation version; Version B of the item introduced a minimal verbal load in being the verbal translation of that mathematical operation; and Version C of the item represented the maximal verbal load in embedding that operation within a narrative context. (See Appendix A for examples.) To construct these items, experimental base items were gathered from any of the three versions and then modified into the remaining two versions. The sequence of the task and computational format remained the same in all three versions, and no irrelevant data or verbiage was introduced into any version.

To gain a better understanding of male versus female achievement in mathematics, Marshall suggests that performance be compared on "well-defined types of items" in order to determine whether the sexes differ (Marshall, 1984). The present study attempts to do this in a more carefully controlled way than encountered in previous research. Doolittle's classification scheme,

for example, blurs the distinction between the present study's Version B and Version C items (Doolittle, 1987). Caldwell & Golden's classification scheme (Caldwell & Golden, 1979), on the other hand, fails to account for the present study's Version A items. In neither case are the items systematically controlled and varied through the three versions from the pure computational to the maximum verbal presentation of the same item.

The experimental items were assembled into tryout units in such a fashion that only one version of any of the twenty-one experimental items appeared in a single tryout unit. Each tryout unit was randomly assembled to include A, B and C versions of items so that no examinee took a tryout unit of only one item format. Each experimental unit was administered to approximately 3000 examinees (in randomly equivalent samples such that only one tryout unit was administered to any one examinee) as part of the ACT Assessment tryout procedures.

To confirm whether verbal factors affect performance, the data were initially analyzed by using item analyses from each of the three versions. The difficulty (percentage of examinees answering an item correctly) and discrimination values (biserial correlation) of each version were studied and compared.

Because it seemed likely that Version C items might require a different skill, or additional skills, than Version A items, the difference between the proportion of correct responses for low- and high-ability examinees for each version of an item was studied by means of a three-way analysis of variance (one gender factor, one item version factor, and one ability factor).

Another analysis, the standardization technique (Dorans, 1986), was performed to examine if gender differences increased as the degree of verbal loading increased. For an item, the standardization statistic, D_{STD} , is computed using the formula

$$D_{STD} = \sum_{s=1}^S \frac{K_s}{T_f} [P_{fs} - P_{bs}]$$

where

P_{fs} is the proportion correct in the focal group (females) at score level s ,

P_{bs} is the proportion correct in the base group (males) at score level s ,

K_s is the relative frequency of focal group examinees at score level s ,

and

T_f is the total number of focal group examinees.

Thus, D_{STD} can be interpreted as the weighted average proportion correct difference between males and females for a given item. The conditioning score used in this study was the raw score that examinees obtained on the 40-item ACT Mathematics Usage Test. Positive D_{STD} values would indicate items on which female examinees outperformed male examinees, whereas negative values would indicate items on which males outperformed females.

RESULTS

Preliminary item analysis results concerning difficulty and discrimination confirm the hypothesis that verbal loading affects performance. Version A items, with a mean p-value of .57, were easiest overall; Version C items, with a mean p-value of .49, were the most difficult; Versions B (mean p-value of .51) and C items were most similar in difficulty. A biserial was used to measure the discrimination of each item. Version A items were most discriminating, and Version C items were least

discriminating. The standard deviations for the three versions for both difficulty and discrimination were very similar, ranging from .22 to .16.

The examinee population was then divided into male and female groups and proportion correct and point biserial correlations were computed for each sex. Each gender group was further divided into three ability groups: Low, Medium, and High. Regarding male versus female performance, one might assume that, since verbal ability is considered important in the learning of mathematics (Aiken, 1971b), and since females are believed to generally excel in this ability (Maccoby & Jacklin, 1974), females would excel in Version C items. This, however, was not the case. The mean item difficulties across the three versions by sex and by ability level are summarized in Table 1. Version A items were slightly easier for females than males across all three ability levels. Versions B and C were slightly easier for males across all three ability levels. The difference between low- and high-ability examinees decreased slightly as the verbal load in the items increased.

Insert Table 1 about here

Point biserial correlations were computed using the examinees' dichotomous item score 1 (correct), 0 (incorrect) and the total score on the 40-item ACT Mathematics Usage Test. These values are displayed in Table 2. For the most part, there is very little difference between the point biserial for males and females within item type. However, some noticeable differences do exist; for example, Type C of item 1 ($r_M = .41$ vs. $r_F = .27$); Type A of item 3 ($r_M = .39$ vs. $r_F = .28$); and Type B of item 16 ($r_M = .23$ vs. $r_F = .13$). These, however, appear to be just random occurrences. When comparing point biserials across item type, no clear pattern emerges. Within gender

group, the correlations tend to decrease as the verbal load increases, although there are several exceptions.

Insert Table 2 about here

The results of the analysis of variance are summarized in Table 3. No significant interactions were found. The item type main effect and the ability level main effect were both found to be statistically significant. Since the formation of the ability levels is based on an examinee's overall performance on the ACT Mathematics Usage Test, it is not surprising that the ability level main effect was significant. The main effect for item type does indicate that Version C items are significantly more difficult for the entire population than Version A items.

Insert Table 3 about here

Results of the standardization analysis are shown in Table 4. A priori it was established that $|D_{STD}| \geq .1$ would be considered the level at which one group had significantly outperformed the other group. Out of the twenty-one items, two were found to significantly favor males (item 12, Version B, and item 21, Version C) and one (item 9, Version A) was found to significantly favor females.

Insert Table 4 about here

However, if the degree of significance is ignored, a general trend can be seen. As the verbal load increased, the number of items on which males

outperformed females also increased. That is, out of 21 items males outperformed females on only seven Type A items, on ten Type B items, and on fifteen Type C items.

DISCUSSION

Results of this study suggest that increasing the verbal load of an item does affect item performance. Perhaps this occurrence further suggests that the item is becoming more multidimensional. That is, within the context of this study, increasing the verbal load of mathematics items may necessitate the use of verbal and reading skills as well as mathematics skills. Regarding gender differences, the results of both the analysis of variance and the standardization procedure indicate that there is little difference in performance between males and females on each of the three types of items studied here.

Given that critical decisions are often made on the basis of test scores, it would behoove a test developer to be able to provide an unequivocal explanation of the skill being tested in a particular mathematics item format. Those who are responsible for constructing test forms would be helped by knowing that performance among different groups might vary according to item format: they could avoid constructing test forms that systematically result in differential performance disadvantageous to a particular group unless there are curriculum-related reasons to do so. To the extent that a low score on a mathematics test represents difficulties with verbal/reading skills rather than with mathematics, then the validity of the test is diminished. Perhaps this study sheds some light on the issue at least to the extent of pointing the direction for future work--namely, to examine the relationship between multidimensionality and verbal loading, and to examine the relationship between reading ability and verbal loading.

BIBLIOGRAPHY

- Aiken, L. R. (1971a). Intellectual variables and mathematics achievement: Directions for research. Journal of School Psychology, 9, 201-212.
- Aiken, L. R. (1971b). Verbal factors in mathematics learning: A review of research. Journal for Research in Mathematics Education, 1971, 2, 304-313.
- Benbow, C. P., & Stanley, J. C. (1982). Consequences in high school and college of sex differences in mathematical reasoning ability: A longitudinal perspective. American Educational Research Journal, 19, 598-622.
- Caldwell, J. H., & Golden, G. A. (1979). Variables affecting word problem difficulty in elementary school mathematics. Journal for Research in Mathematics Education, 10, 323-336.
- Doolittle, A. D., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. Journal of Educational Measurement, 24, 157-166.
- Dorans, Neil J. (1986). Two approaches to assessing unexpected differential item performance: Standardization and the Mantel-Haenszel method. A paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Maccoby, E. E., & Jacklin, C. N. (1974). Psychology of sex differences. Palo Alto, California: Stanford University Press.
- Marshall, S. P. (1984). Sex differences in children's mathematics achievement: Solving computations and story problems. Journal of Educational Psychology, 76, 194-204.
- Muscio, R. D. (1962). Factors related to quantitative understanding in the sixth grade. Arithmetic Teacher, 9, 258-262.

Paul, J., Nibbelink, W. H., & Hoover, H. D. (1986). The effects of adjusting readability on the difficulty of mathematics story problems. Journal for Research in Mathematics Education, 17, 163-171.

Table 1
P-Value Means and Standard Deviations for Item Types
by Sex by Ability Level

Ability	Group	Version A		Version B		Version C	
		$\bar{X}$	STD	$\bar{X}$	STD	$\bar{X}$	STD
Low	Male	.31	.15	.30	.15	.30	.12
	Female	.32	.16	.29	.17	.29	.12
Mid	Male	.55	.17	.50	.16	.47	.16
	Female	.57	.18	.47	.16	.43	.16
High	Male	.81	.13	.78	.11	.74	.12
	Female	.82	.13	.76	.13	.70	.13

Table 2
Point Biserial Correlations

Item	Version A		Version B		Version C	
	M	F	M	F	M	F
1	.48	.43	.48	.39	.41	.27
2	.48	.43	.49	.44	.28	.26
3	.39	.28	.26	.25	.18	.16
4	.42	.41	.41	.37	.40	.45
5	.57	.55	.46	.38	.40	.40
6	.34	.36	.30	.31	.34	.30
7	.50	.48	.44	.45	.39	.35
8	.52	.49	.57	.54	.51	.48
9	.41	.39	.48	.44	.36	.31
10	.45	.46	.47	.48	.46	.47
11	.47	.51	.48	.48	.35	.29
12	.40	.47	.42	.40	.28	.27
13	.33	.31	.35	.35	.38	.40
14	.47	.49	.40	.40	.30	.32
15	.34	.32	.31	.34	.34	.32
16	.16	.14	.23	.13	.40	.42
17	.51	.51	.46	.47	.48	.37
18	.31	.34	.43	.37	.45	.43
19	.39	.38	.40	.40	.42	.35
20	.37	.36	.38	.34	.33	.27
<u>21</u>	<u>.52</u>	<u>.53</u>	<u>.44</u>	<u>.41</u>	<u>.41</u>	<u>.41</u>
Mean:	.42	.41	.41	.39	.37	.35

Table 3

ANOVA Summary Table: Item Version by Sex by Ability Level

Source	SS	df	MS	F	F prob.
Item Version	396,880	2	198,439.78	9.26	.00
Sex	13,540	1	13,539.85	.63	.43
Ability Level	13,701,213	2	6,850,606.27	319.55	.00
Item Version x Sex	31,709	2	15,854.67	.74	.48
Item Version x Ability	96,906	4	24,226.49	1.13	.34
Sex x Ability Level	131,024	8	16,378.03	.76	.64
Sex x Ability Level x Item Version	6,581	4	1,645.18	.08	.64
Total	21,952,745	376	58,384.96		

Table 4
Standardization Values

Item	A	<u>Version</u> B	C
1	.03	.01	.06
2	.03	-.03	-.01
3	.02	-.01	.00
4	.04	.04	-.02
5	.07	-.09	-.05
6	.00	.00	-.03
7	.07	.02	-.03
8	.06	.02	-.04
9	.10	.00	-.05
10	.03	.00	.03
11	.04	-.06	-.07
12	-.05	-.12	-.06
13	.03	-.01	.00
14	.05	-.03	.00
15	-.01	.01	-.01
16	.03	.03	-.07
17	-.02	.02	-.06
18	-.04	.01	-.05
19	-.04	-.01	.02
20	-.03	-.07	-.03
21	-.04	-.10	-.05

APPENDIX A

Version A Item

2. If $.30x - .10x = 240$, then $x = ?$

- F. 2,400
- G. 1,200
- H. 800
- J. 600
- K. 300

Version B Item

2. The difference between 30% of a number and 10% of that number is 240. What is the original number?

- F. 2,400
- G. 1,200
- H. 800
- J. 600
- K. 300

Version C Item

2. Last month, an organization spent 30% of its income on postage and 10% of its income on the telephone bill. The postage costs were \$240 more than the telephone bill. What was the organization's income last month?

- F. \$2,400
- G. \$1,200
- H. \$ 800
- J. \$ 600
- K. \$ 300